MAPSI — cours 6 : Chaîne de Markov

Vincent Guigue, Thierry Artières
vincent.guigue@lip6.fr

LIP6 – Université Paris 6, France

Introduction

- Les problèmes traités jusqu'ici :

$$X = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1d} \\ \vdots & & & \\ x_{N1} & x_{N2} & \cdots & x_{Nd} \end{bmatrix}, \text{ et parfois : } Y = \begin{bmatrix} y_1 \\ \vdots \\ y_N \end{bmatrix}$$

- Chaque individu $\mathbf{x} = [x_1, x_2, \cdots, x_d]$ est un vecteur
- **Les séquences** ne rentrent pas dans ce cadre

Traitement des séquences

Tâches :

- Classification / Clustering
- Etiquetage / Segmentation
- Génération de séquences

- Reconnaissance de chaine de caractères

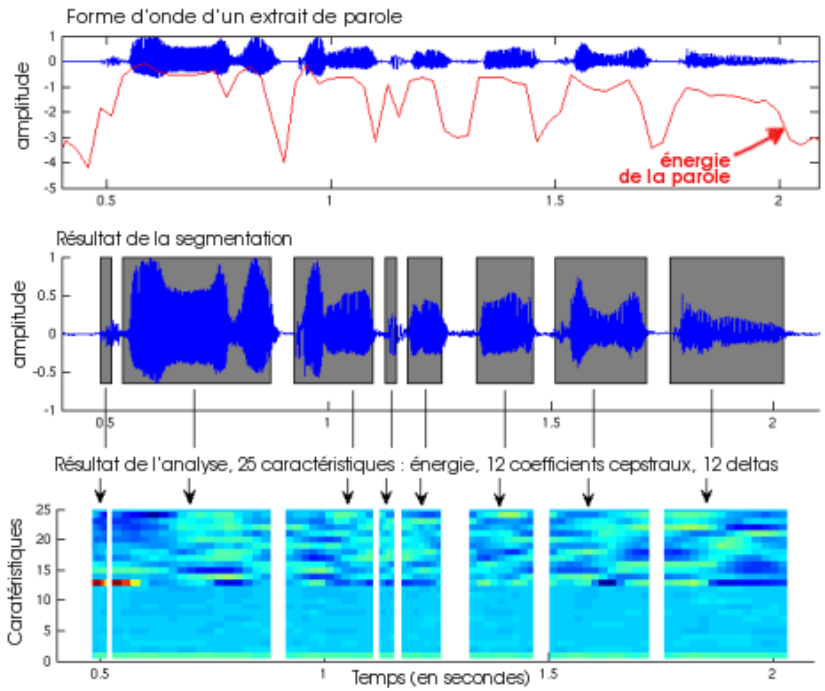
- Reconnaissance de paroles
- Génération/reconnaissance de mouvements
- Reconnaissance de mouvements (2)

Traitement des séquences

Tâches :

- Classification / Clustering
- Etiquetage / Segmentation
- Génération de séquences

- Reconnaissance de chaine de caractères
- Reconnaissance de paroles

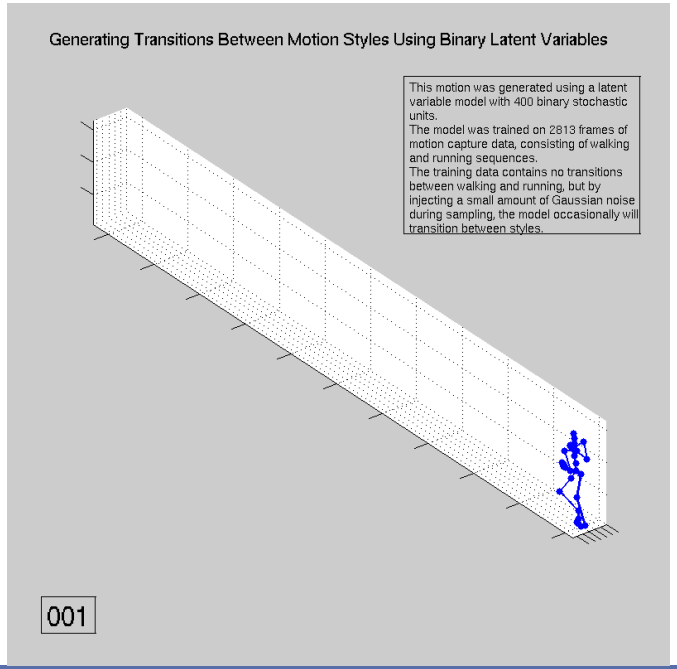


Traitement des séquences

Tâches :

- Classification / Clustering
- Etiquetage / Segmentation
- Génération de séquences

- Reconnaissance de chaine de caractères
- Reconnaissance de paroles
- Génération/reconnaissance de mouvements

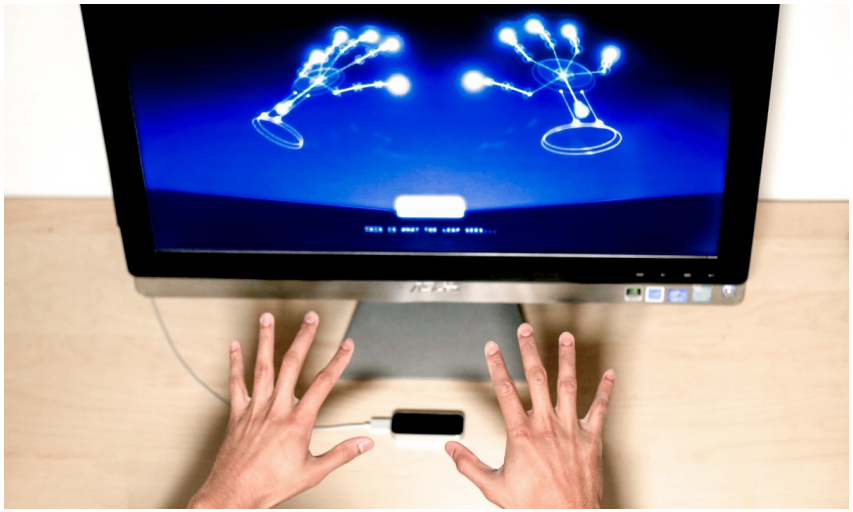


Traitement des séquences

Tâches :

- Classification / Clustering
- Etiquetage / Segmentation
- Génération de séquences

- Reconnaissance de chaine de caractères
- Reconnaissance de paroles
- Génération/reconnaissance de mouvements
- Reconnaissance de mouvements (2)



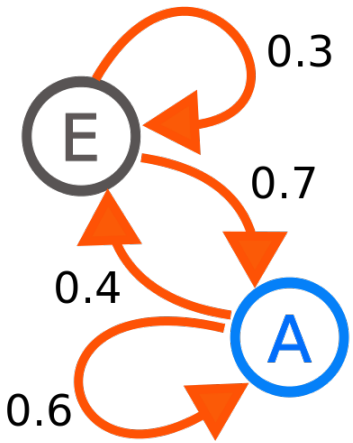
Approche générative

Problème

- Difficile d'étendre les méthodes standards de classification ou de clustering à des données de taille variable
- Mais plus facile de concevoir des modèles génératifs de données de taille variable
- Approche de classification par apprentissage des densités

Approche vectorielle : $X = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1d} \\ \vdots & & & \\ x_{N1} & x_{N2} & \cdots & x_{Nd} \end{bmatrix}$

Approche séquentielle :



Modèles génératifs

Vincent Guigue, Thierry Artières
vincent.guigue@lip6.fr

LIP6 – Université Paris 6, France

Rappel sur les modèles génératifs

1. Choix d'une modélisation des données : $p(\mathbf{x}|\theta)$
2. Apprentissage = trouver θ
3. Application possible : décision bayesienne

$$r(\mathbf{x}) = \arg\max_k p(\theta_k|\mathbf{x}) = \frac{p(\mathbf{x}|\theta_k)p(\theta_k)}{p(\mathbf{x})}$$

4. Application bis : génération de $\tilde{\mathbf{x}} \sim \mathcal{D}(\theta_k)$

Apprentissage d'un modèle génératif $\Leftrightarrow$ Estimation de densité

- Estimer θ_k = estimer une densité de probabilité d'une classe
- Hypothèse (forte) : les θ_k sont supposés indépendants
- Techniques d'estimation des θ_k

Maximum de vraisemblance

- $D = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$ exemples supposés générés par $p(\mathbf{x}|\theta)$
- Adéquation entre les données et le modèle
  - Notion de vraisemblance des observations
  - Hyp : les données sont indépendantes

$$\mathcal{L}(D, \theta) = p(D|\theta) = \prod_{i=1}^{N} p(\mathbf{x}_i|\theta)$$

- Optimisation :

$$\theta^\star = \arg\max_\theta(\mathcal{L}(D, \theta)) = \arg\max_\theta(\log\mathcal{L}(D, \theta))$$

- Résolution :
  - Analytique : $\frac{\partial\mathcal{L}(D, \theta)}{\partial\theta} = 0$
  - Approchée : EM, gradient...

Traitement des séquences

- Prise en compte de l'information séquentielle présente dans les données
- Nombreux domaines
  - Séries temporelles : finance, consommation, marketing, etc
  - Parole, biologie, langue,
- Méthodes & problématiques
  - Prévision de séries : AR, ARMA, etc
  - Matching (et classification) : Dynamic time warping
  - Modèles génératifs : MMC, réseaux de neurones, etc
- Présentation
  - Modèles de Markov, Modèles de Markov Cachés

Chaînes de Markov

- Outil pour faire de la prévision dans des espaces discrets
- Chaîne de Markov d'ordre k
  - Séquence de variables aléatoires $S = (s_1, \ldots, s_T)$ qui prend ses valeurs dans un ensemble fini d'états $Q = (q_1, ..., q_N)$ et qui vérifie les propriétés dites de Markov :
    - Horizon de taille k :

$$p(s_{t+1} = q_j | s_1, \ldots, s_t) = p(s_{t+1} = q_j | s_{t-k+1}, \ldots, s_t)$$

    - Stationnarité :

$$p(s_{t+1} = q_j | s_1, \ldots, s_t) = p(s_{k+1} = q_j | s_1, \ldots, s_k)$$

  - Pour simplifier les notations, on se limite dans la suite à des chaînes d'ordre 1.
- Exemple : météo sur un an (soleil, nuage, pluie)
  - $Q = [So, Nu, Pl]$
  - $S = [s_1 = Nu, s_2 = So, \ldots, s_{365} = Pl]$

CM d'ordre 1

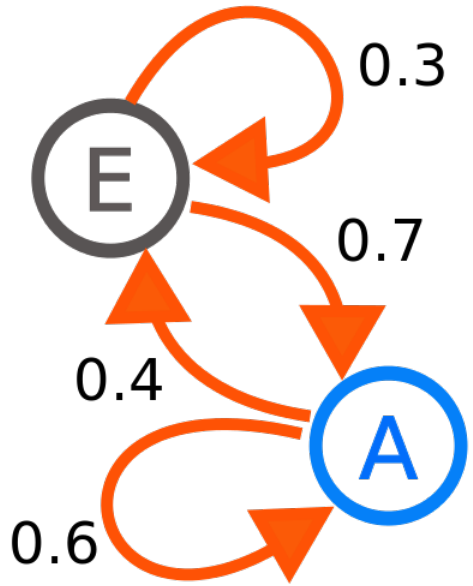
- Une chaîne de Markov d'ordre 1 est entièrement spécifiée par la donnée :

- d'une matrice de transition

$$A = \left[a_{ij} = p(s_{t+1} = q_j | s_t = q_i)\right]$$

- et des probabilités initiales :

$$\Pi = [\pi_i = p(s_1 = q_i)]$$



- Probabilité d'une séquence

$$p(S|\lambda) \quad = \quad p(s_1, \ldots, s_T|\lambda)$$

Hypothèse markovienne d'ordre 1

Décomposition du calcul

$$
\begin{aligned}
p(S|\lambda) &= p(s_1, \ldots, s_T|\lambda) \\
&= p(s_T|s_1, \ldots, s_{T-1}, \lambda) \times p(s_1, \ldots, s_{T-1}|\lambda) \\
&= p(s_T|s_1, \ldots, s_{T-1}, \lambda) \times p(s_{T-1}|s_1, \ldots, s_{T-2}, \lambda) \ldots \\
&\quad \times p(s_1, \ldots, s_{T-2}|\lambda)
\end{aligned}
$$

$$
= \prod_{t=2}^{T} p(s_t|s_1, \ldots, s_{t-1}, \lambda) p(s_1|\lambda)
$$

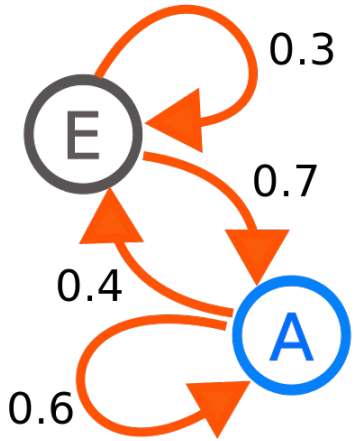
Après hypothèse d'ordre 1 :

$$
p(S|\lambda) = \prod_{t=2}^{T} p(s_t|s_1, \ldots, s_{t-1}, \lambda) p(s_1|\lambda) = \prod_{t=2}^{T} p(s_t|s_{t-1}, \lambda) p(s_1|\lambda)
$$

$$
= \pi_{s_1} \prod_{t=2}^{T} a_{s_{t-1}, s_t}
$$

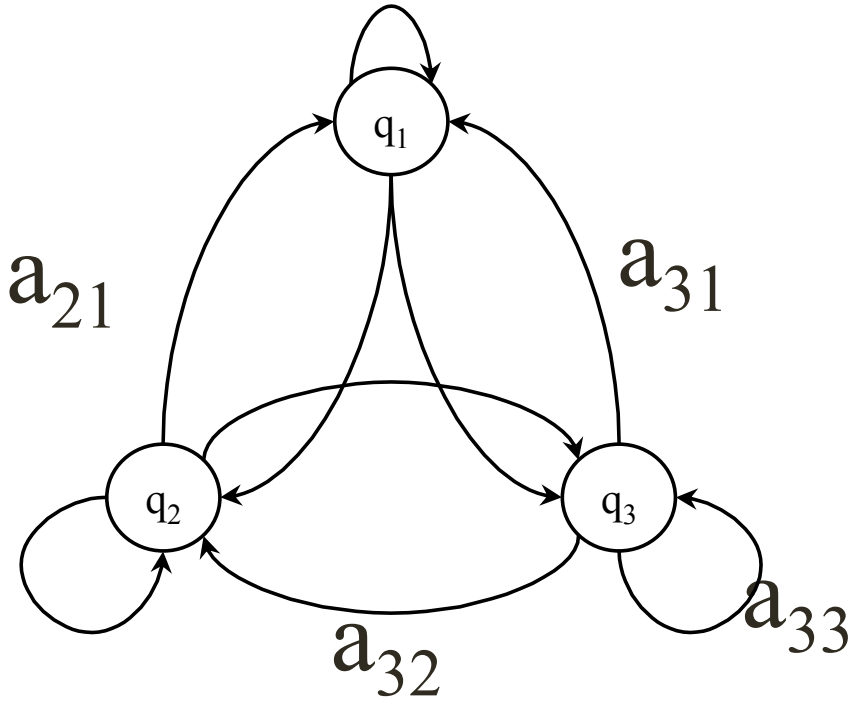
Représentation graphique

Automate basique :



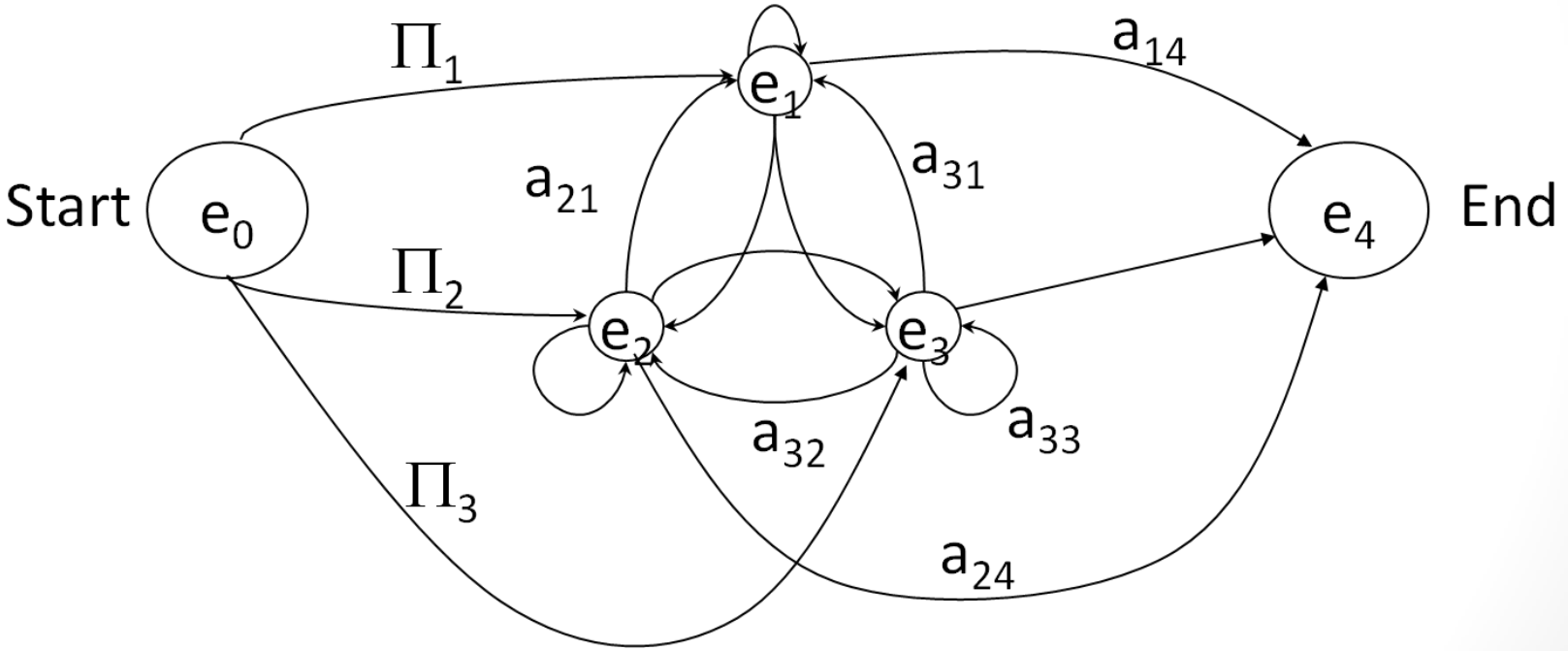
Représentation graphique

En introduisant les notations a, π



Représentation graphique

Avec des noeuds identifiés de début/fin



Algorithme génaratif

Algorithm 1: Génération d'une séquence S

Data: A, Π
Result: S
$S \leftarrow []$;
Tirer s_1 en fonction de Π;
$s_t \leftarrow s_1$, $t \leftarrow 1$;
$S \leftarrow [S,\ s_{courant}]$;
while s_t *n'est pas l'état final* **do**
    $s_{t+1} \leftarrow$tirage selon $(A(s_t, :))$;
    $t \leftarrow t + 1$;

- Plusieurs variantes dans la clause du *while*
- Comment effectuer un tirage selon une loi de probabilité discrète ?

Outil pour le tirage aléatoire selon une loi discrète

Soit la loi :

A	1	2	3
$P(A)$	0.3	0.2	0.5

Comment effectuer un tirage selon $P(A)$?

1. Faire la somme cumulée de la loi

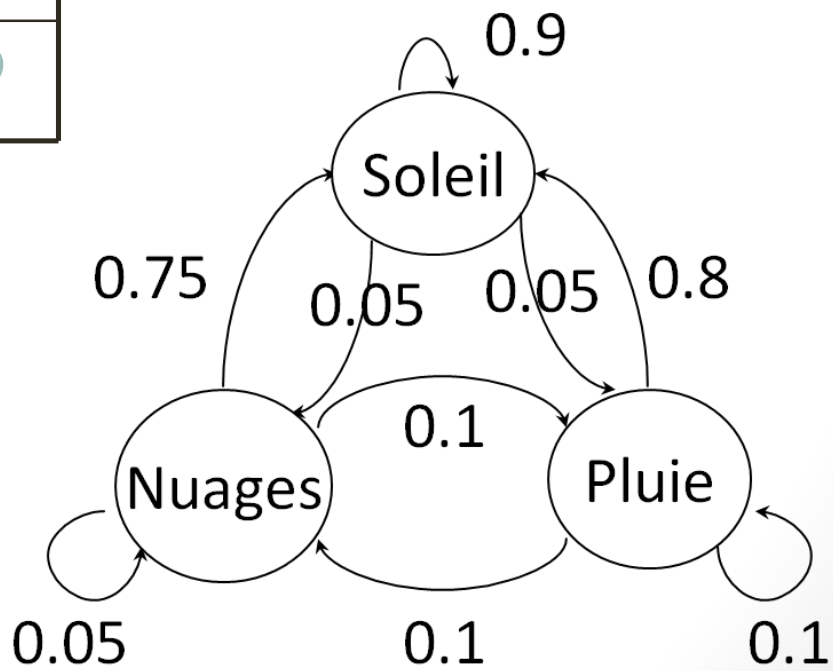
A	1	2	3
cumsum	0.3	0.5	1

2. Tirer un nombre x entre 0 et 1 selon la loi uniforme
3. Initialiser $vx = 1$
4. Tant que $cumsum[vx] < x$
   - $vx++$

Génération (sur un exemple)

- q_1 = Pluie, q_2 = Nuages, q_3 = Soleil
- A =

0.1	0.1	0.8
0.1	0.15	0.75
0.05	0.05	0.9



PSSSSSSSSSSSSNSSSSSSSSSSSSNSSSSSSSNSSSSNSS

SSSSSSSSSSSSSSPSSSSSSNPSSSSSPNSSSSSSSNSSS

SSSSPSSSSSSSSSSPSSSSSSSSSSSSSSSSPSSSSSSSSS

Problèmes alternatifs

- Quelle est la probabilité d'observer une séquence de soleil de longueur d ?
- Quelle est la durée moyenne d'une séquence consécutive de soleil ?

Quelle est la probabilité d'observer une séquence de soleil de longueur d ?
Loi géométrique
Notons la longueur de la sous-séquence de soleil D_S,

$$P(D_S = d) = a_{ss}^{d-1}(1 - a_{ss})$$

Problèmes alternatifs

- Quelle est la probabilité d'observer une séquence de soleil de longueur d ?
- Quelle est la durée moyenne d'une séquence consécutive de soleil ?

Quelle est la longueur moyenne d'une séquence de soleil ?
Espérance de la loi géométrique

$$E[D_S] = \sum_{d=1}^{\infty} d\ a_{ss}^{d-1}(1 - a_{ss}) = \frac{1}{1 - a_{ss}}$$

Problèmes alternatifs

- Quelle est la probabilité d'observer une séquence de soleil de longueur d ?
- Quelle est la durée moyenne d'une séquence consécutive de soleil ?

Quelle est la longueur moyenne d'une séquence de soleil ?
Espérance de la loi géométrique

$$E[D_S] = \sum_{d=1}^{\infty} d\ a_{ss}^{d-1}(1 - a_{ss}) = \frac{1}{1 - a_{ss}}$$

Sketch of proof (wikipedia) avec $k = d - 1$ et $p = 1 - a_{ss}$:

$$\begin{aligned}
\mathrm{E}(Y) &= \textstyle\sum_{k=0}^{\infty}(1-p)^k p \cdot k \\
&= p\textstyle\sum_{k=0}^{\infty}(1-p)^k k \\
&= p(1-p)\left[\frac{d}{dp}\left(-\textstyle\sum_{k=0}^{\infty}(1-p)^k\right)\right] \\
&= -p(1-p)\frac{d}{dp}\frac{1}{p} = \frac{1-p}{p}.
\end{aligned}$$

Problèmes alternatifs (2)

- Il fait soleil...
- Quel temps fera-t-il dans N jours ? (distribution de probabilités)

Problèmes alternatifs (2)

- Il fait soleil...
- Quel temps fera-t-il dans N jours ? (distribution de probabilités)

1. Je rentre sur la ligne *soleil*... $s_0 = S$
2. A $t = 1$, $\{a_{s\cdot}\}$ me donne la distribution des probabilités des états
$$p(s_1 = q_i) = a_{S,i}$$
3. ATTENTION : Ensuite il s'agit d'un treillis

$$p(s_2 = q_j) = \sum_i p(s_1 = q_i) p(s_2 = q_j | s_1 = q_i)$$

Problèmes alternatifs (2)

- Il fait soleil...
- Quel temps fera-t-il dans N jours ? (distribution de probabilités)

1. Je rentre sur la ligne *soleil*... $s_0 = S$
2. A $t = 1$, $\{a_{s\cdot}\}$ me donne la distribution des probabilités des états

$$p(s_1 = q_i) = a_{S,i}$$

3. ATTENTION : Ensuite il s'agit d'un treillis

$$p(s_2 = q_j) = \sum_i p(s_1 = q_i) p(s_2 = q_j | s_1 = q_i)$$

Ecriture matricielle simple :

$$p(s_N | s_0 = S) = a_{S,\cdot} \times A^{N-1}$$

Problèmes alternatifs (3) - et propriétés

- **Stationnarité** : existe-t-il une mesure stationnaire μ telle que $\mu = \mu A$?
  - μ = pondération stationnaire (inchangée après une transition)
  - si $\forall i,\ \mu_i \geq 0,\ \sum_i \mu_i = 1$: μ est alors une distribution stationnaire
  - si A est irreductible, μ est unique et : $\mu =$ distribution moyenne des états

CM irreducible finie

Les chaines sur lesquelles nous travaillons sont irréductibles et finies : partant de chaque état, on y revient en un nombre moyen d'étapes fini.

Problèmes alternatifs (4) - et propriétés

- Des sous-séries d'observations sont-elles récurrentes dans une CM ?

Périodicité

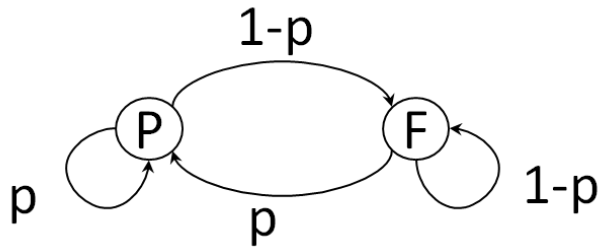
- Un état est dit périodique de période k ($k > 1$), si on ne peut y revenir (après l'avoir quitté) qu'en un nombre d'étapes multiples de k.
- La période d'une CM est définie comme le PGCD de la période de tous ses états.
- La période d'une CM est égale au PGCD de la longueur de tous les circuits (élémentaires) du graphe associé.
- Une CM est dite *apériodique* si sa période est égale à 1.

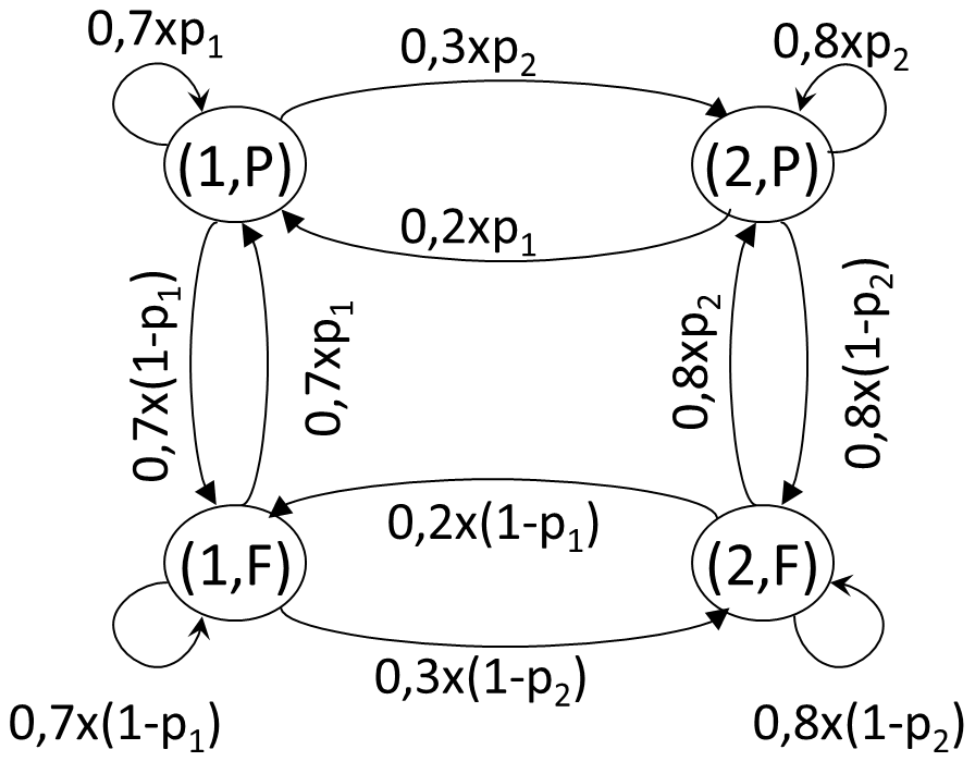
Modélisation (2)

- Séquence de lancers de pièce(s)... Mais combien y en a-t-il ?

p_k : probabilité de faire *pile* avec la pièce *k*

p : probabilité de faire *pile*





Modélisation (3)

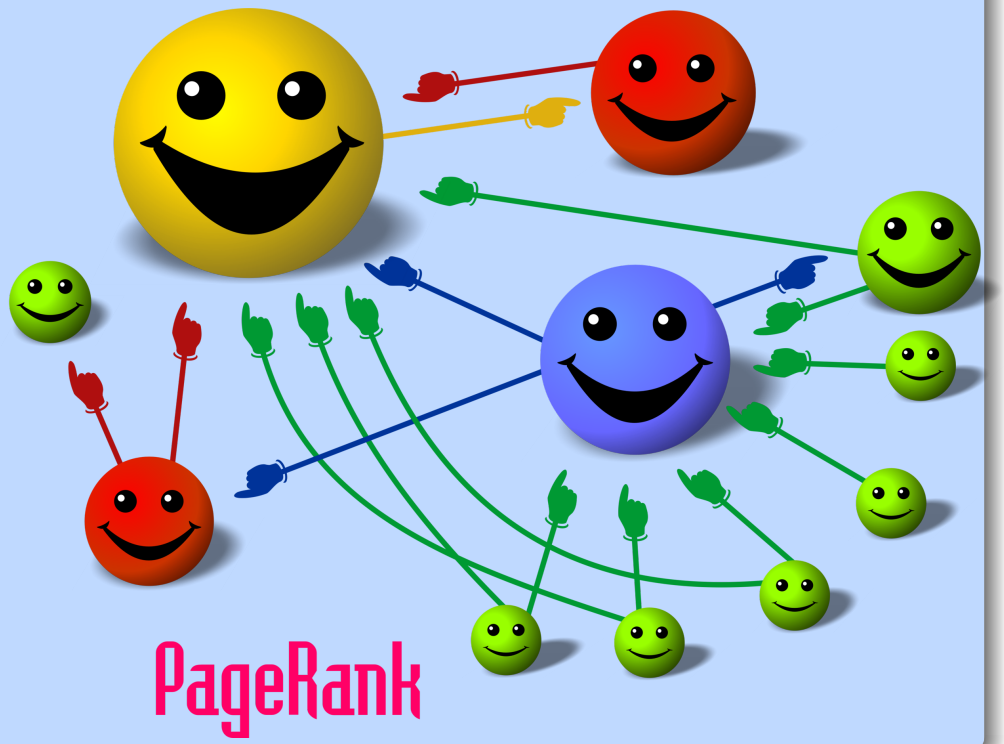
- Modélisation de parcours utilisateur sur un site web
  - Catégorsation / publicité personnalisée
  - Optimisation du site / pré-chargement de pages
- 1 trace = longueur variable...

Modélisation (3)

- Modélisation de parcours utilisateur sur un site web
  - Catégorsation / publicité personnalisée
  - Optimisation du site / pré-chargement de pages
- 1 trace = longueur variable...

Modèle

- **Etats** :
  page du site
- **Transitions** :
  hyperliens



Modélisation (4)

Modèles de N-grams

- Construire un modèle de langage qui permette de capturer la succession des mots
- Modèle de N-grams = CM d'ordre $N-1$
- Exemple : vocabulaire 20K mots

Modèle	Nb paramètres
Bigram	$20k \times 19k = 4\ 10^8$
Trigram	$8\ 10^{12}$
4-gram	$1.6\ 10^{17}$

Modelisation (4) suite

- $\Rightarrow$ En général, nous nous limitons aux N-grams dont le nombre d'occurence dépassent un certain seuil (de nombreuses combinaisons d'existent pas !)
- $p(w_j|w_{j-N},\ldots,w_{j-1}) = \dfrac{p(w_{j-N},\ldots,w_j)}{p(w_{j-N},\ldots,w_{j-1})}$
- En pratique, besoin d'estimateurs plus robustes
  - eg : modèle d'interpolation de Jelinek :

$$p(w_j|w_{j-2},w_{j-1}) = \lambda_1 p(w_j) + \lambda_2 p(w_j|w_{j-1}) + \lambda_3 p(w_j|w_{j-2},w_{j-1})$$

$$\lambda_1 + \lambda_2 + \lambda_3 = 1$$

Jelinek 1997 - Statistical Methods for Speech Recognition

Apprentissage des chaines de Markov

Vincent Guigue, Thierry Artières
vincent.guigue@lip6.fr

LIP6 – Université Paris 6, France

Questions

- Etant donnée une séquence d'états, calculer sa probabilité
  Vu précédemment :

$$
\begin{aligned}
p(S|\lambda) &= \prod_{t=2}^{T} p(s_t|s_1,\ldots,s_{t-1},\lambda)p(s_1|\lambda) = \prod_{t=2}^{T} p(s_t|s_{t-1},\lambda)p(s_1|\lambda) \\
&= \pi_{s_1} \prod_{t=2}^{T} a_{s_{t-1},s_t}
\end{aligned}
$$

- Comment apprendre une CM à partir d'exemples ?
- Comment faire de la classification de séquences avec des CM ?

Apprentissage

- Soit une base de séquences $B = \{S^1, \ldots, S^K\}$ (N états possibles)
- Critère de vraisemblance

$$\log \mathcal{L}(B, \lambda) = \log(\prod_{k=1}^{N} p(S^k|\lambda)) = \sum_{k} \log(p(S^k|\lambda))$$

- Optimisation :

$$\lambda^\star = \arg\max_{\lambda} \log \mathcal{L}(B, \lambda)$$

- Contraintes :

$$\forall i \in [1,\ N], \sum_{j=1}^{N} a_{ij} \quad = \quad 1$$

$$\sum_{j=1}^{N} \pi_j = 1$$

Optimisation Lagrangienne

- Critère intégrant les contraintes :

$$\mathcal{C}(\lambda) = \mathcal{L}(B, \lambda) - \sum_{i=1}^{N} \nu_i \left(\sum_{j=1}^{N} a_{ij} - 1 \right) - \eta \left(\sum_{j=1}^{N} pi_j - 1 \right)$$

- Si la dérivée par rapport au coefficient de contrainte est nulle, la contrainte est satisfaite :

$$\frac{\partial \mathcal{C}(\lambda)}{\partial \eta} = 0 \Leftrightarrow \sum_{j=1}^{N} pi_j - 1 = 0$$

- Résolution au tableau... Et optimum en :

$$a_{ij} = \frac{n_{ij}}{n_{i.}} \quad \pi_j = \frac{l_j}{K}, \text{ avec : } n_{i.} = \sum_{j} n_{ij}$$

En pratique

- Approche par comptage
- Calcul des fréquences des évènements = solution au sens MV
- Chaque ligne de A est une distribution (sommant à 1)
- En faisant une hypothèse de stationnarité, il est possible d'estimer les π_j sur toute la base de données :

$$\text{classique : } \pi_j = \frac{l_j}{K} \quad \text{alternative stationnaire : } \pi_j = \frac{n.j}{\sum_{ij} n_{ij}}$$

Distance entre séquences

Vincent Guigue, Thierry Artières
vincent.guigue@lip6.fr

LIP6 – Université Paris 6, France

Idée générale

- Similarité/distance = outil de base
  - k-plus proches voisins...
- La similarité peut concerner une partie seulement du signal
  - Reflexion sur les besoins spécifiques

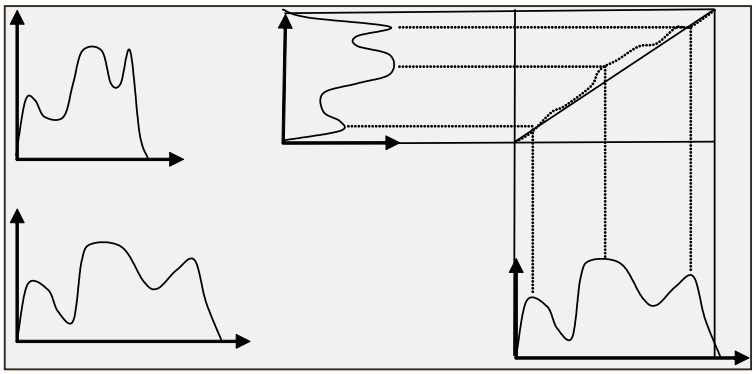
Dynamic Time Warping (DTW)

- L'analyse du signal est locale (cf. stationnarité)
- Unités de reconnaissance plus globales (phonèmes, mots, ...)
- $\Rightarrow$ Nécessité de comparer des séquences de vecteurs
- **DTW = distance entre séquences**
  - ayant des longueurs différentes
  - insensible à certaines variabilités d'élocution
  - calculable efficacement

Distance entre séquences

Idée :

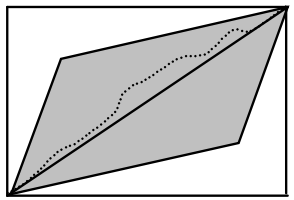
- Existence d'une distance entre séquences capable de
  - Prendre en compte les différences de rythme dans les séquences
  - Comparer des séquences de longueur différente
- Dynamic Time Warping (DTW)

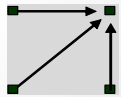


La distance entre les séquences est la somme des distances entre éléments mis en correspondance par l'alignement

Appariement avec contraintes

- ❑ Continuité locale
- ❑ Alignement quasi linéaire
- ❑ Début et fin synchrones





Distance entre séquences (2)

Notion de chemin d'alignement :

- Chemin : $c = \{(i_k, j_k)\}_{k=1,\ldots,K}$ tel que :

$$\forall k,\ (i_k, j_k) = \begin{cases} (i_{k-1} - 1, j_{k-1}) & i_1 = j_1 = 1 \\ (i_{k-1} - 1, j_{k-1} - 1) & j_K = T_2 \\ (i_{k-1}, j_{k-1} - 1) & i_K = T_1 \end{cases}$$

- Distance suivant un alignement :

$$D_c(S_1, S_2) = \sum_{k=1}^{K} d_{c(k)}(S_1[i(k)], S_2[j(k)])$$

- Distance entre 2 séquences

$$D(S_1, S_2) = \min_c D_c(S_1, S_2)$$

Distance entre séquences (3)

Phase avant :

- calcul des $\forall i, j,\ d(S_1[i], S_2[j])$
- sommes cumulées

3	2	5	2	4	2	2
2	2	3	2	1	4	4
2	1	2	2	2	3	4
1	1	2	1	1	3	2
1	1	3	3	3	3	4

2	2					
1	2					

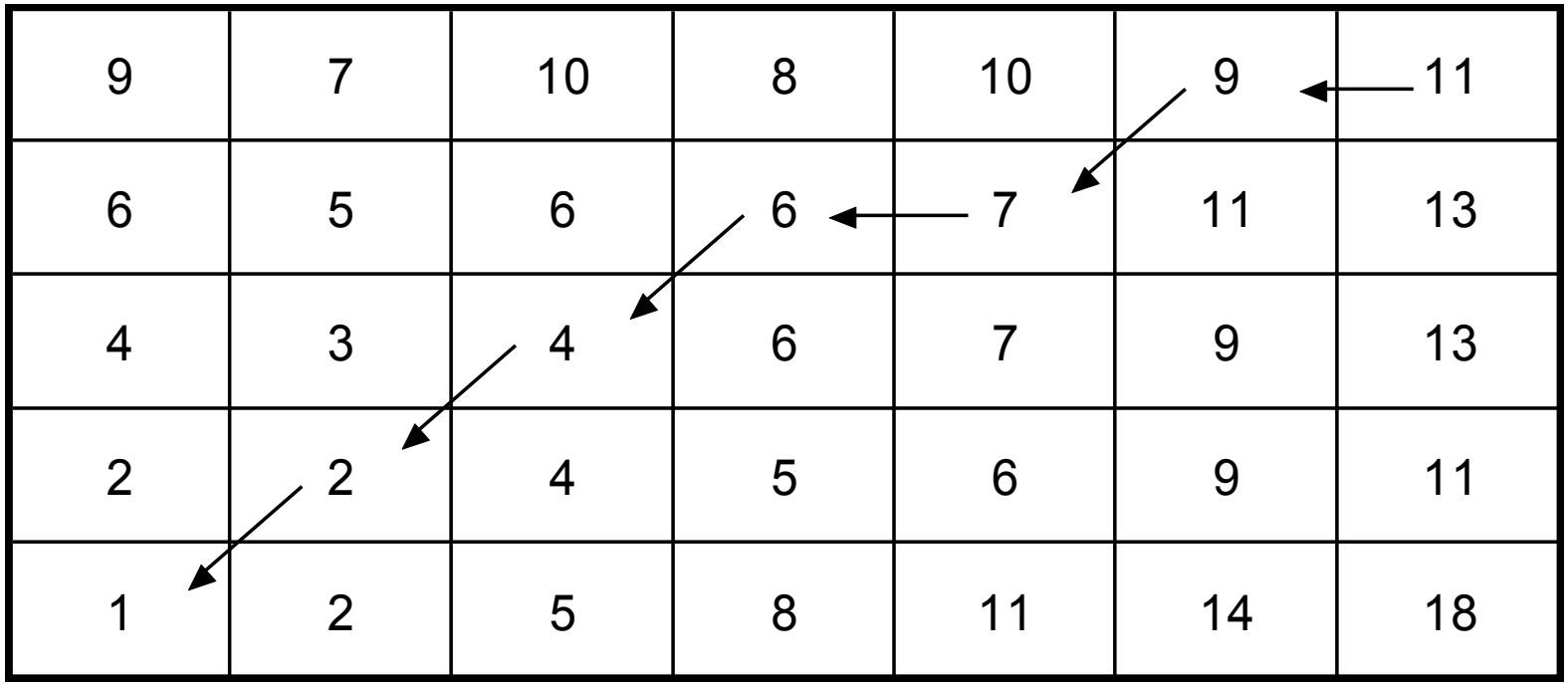
3	3	4				
2	2	4				
1	2	5				

9	7	10	8	10	9	11
6	5	6	6	7	11	13
4	3	4	6	7	9	13
2	2	4	5	6	9	11
1	2	5	8	11	14	18

Distance entre séquences (4)

Phase retour :

- Chemin correspondant au cout minimum



Distance entre séquences (5)

3	2	5	2	4	2	2
2	2	3	2	1	4	4
2	1	2	2	2	3	4
1	1	2	1	1	3	2
1	1	3	3	3	3	4

2	2					
1	2					

3	3	4				
2	2	4				
1	2	5				

9	7	10	8	10	9	11
6	5	6	6	7	11	13
4	3	4	6	7	9	13
2	2	4	5	6	9	11
1	2	5	8	11	14	18

Phase avant

Phase arrière

9	7	10	8	10	9	11
6	5	6	6	7	11	13
4	3	4	6	7	9	13
2	2	4	5	6	9	11
1	2	5	8	11	14	18

Alignement final et distance

x_5	3	2	5	2	4	2	2
x_4	2	2	3	2	1	4	4
x_3	2	1	2	2	2	3	4
x_2	1	1	2	1	1	3	2
x_1	1	1	3	3	3	3	4
	y_1	y_2	y_3	y_4	y_5	y_6	y_7